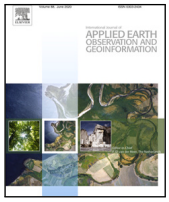
International Journal of Applied Earth Observation and Geoinformation 122 (2023) 103397

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Few-shot object detection on aerial imagery via deep metric learning and knowledge inheritance

Wu-zhou Li [a], Jia-wei Zhou [b], Xiang Li [c], Yi Cao [b], Guang Jin [a,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430022, China
[b] School of Electronic Information, Wuhan University, Wuhan 430022, China
[c] Electrical and Computer Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Keywords:
Few-shot object detection
Aerial imagery
Deep metric learning
Catastrophic forgetting

ABSTRACT

Object detection is crucial in aerial imagery analysis. Previous methods based on convolutional neural networks (CNNs) require large-scale labeled datasets for training to achieve significant success. However, the acquisition and manual annotation of such data is time-consuming and expensive. In this study, we present an original few-shot object detection (FSOD) method that focuses on detecting unseen objects in aerial imagery with limited labeled samples. Specifically, we revisited the multi-similarity network from deep metric learning and incorporated it into a faster region-CNN (R-CNN) architecture for FSOD, learning distinctive feature representations, and effectively improving the performance of unseen class samples. Furthermore, we preserved the knowledge learned from abundant base data by designing a knowledge inheritance module to ease the influence of catastrophic forgetting. We conducted experiments on two benchmark remote sensing image datasets, and the results demonstrated that the proposed methods could achieve a satisfactory performance for FSOD in aerial imagery.

1. Introduction

Object detection means recognizing the category of the objects and marking their locations (Cheng and Han, 2016; Li et al., 2020b). Object detection for aerial images usually focuses on targets such as airplanes, ships, vehicles, bridges, and ports, facilitating many applications such as environmental protection, assistance in rescue, and urban management (Kim et al., 2020; Qiu et al., 2020; Tang et al., 2020; Shi et al., 2019). Modern CNN-based models with their complex and deep structures have made remarkable achievements in object detection for aerial imagery (Cheng and Han, 2016; Li et al., 2020b, 2022; Zhao et al., 2021b; Wolf et al., 2021; Xiao et al., 2021). However, these models often require a massive amount of annotated samples for effective training, otherwise they can suffer from poor generalization and overfitting (Simonyan and Zisserman, 2014). In practice, acquiring sufficient annotated data can be challenging due to the high costs of image collection, expensive human labor, and burdensome manual annotation procedures. Therefore, it is crucial to study object detection for aerial imagery with limited labeled samples.

Few-shot object detectors are designed to learn from novel data with limited samples and base data with sufficient samples to gain the ability to detect objects in both classes (Chen et al., 2018; Liu et al., 2022). Recently, several researches have focused on solving this problem; they can be roughly generalized into two categories: meta-learning-based (Karlinsky et al., 2019; Wang et al., 2019b; Yan et al., 2019; Han et al., 2022; Kang et al., 2019) and transfer-learning-based (Wang et al., 2020; Sun et al., 2021; Fan et al., 2021; Wu et al., 2020). Meta-learning-based methods extract task-agnostic knowledge by simulating and solving a variety of few-shot learning tasks, then use this knowledge to quickly adapt to new tasks with novel data. Another approach applies transfer learning to conventional detection frameworks. For example, the pioneering work two-stage fine-tuning approach (TFA) (Wang et al., 2020) freezes the backbone network, transfers the parameters pretrained using base data and then fine-tunes the last few layers with novel data. Despite its simplicity, the TFA still outperforms the meta-learning-based methods and has attracted much research interest in the field of aerial imagery (Zhao et al., 2021b; Wolf et al., 2021; Xiao et al., 2021).

However, transfer-learning-based models still face two challenges: (1) classifying novel instances into incorrect categories (Sun et al., 2021), and (2) suffering from catastrophic forgetting in the fine-tuning stage (French, 1999). Due to sensor resolutions, illuminations etc., aerial objects are more complex and difficult to detect with limited data, resulting in misclassification of unseen objects and a poor detection performance. Fig. 1 illustrates that the foreground objects are

* Corresponding author.
E-mail address: jinguang5812@sina.com (G. Jin).

https://doi.org/10.1016/j.jag.2023.103397
Received 15 February 2023; Received in revised form 25 May 2023; Accepted 20 June 2023
Available online 3 July 2023

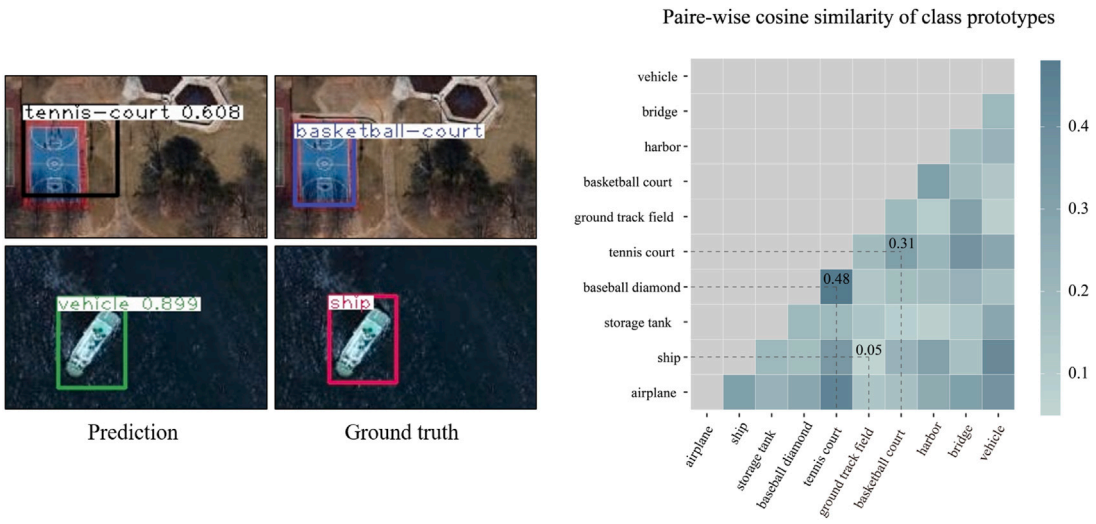


Fig. 1. Left column presents two examples of objects being mislabeled. The right column illustrates the pair-wise cosine similarity between the class prototypes. For instance, the similarity between the ground track field and ship is 0.05 while that between the tennis court and baseball diamond is 0.48. The higher the similarity between objects, the greater the possibility of them being incorrectly classified.

localized accurately but mislabeled into other confusable objects. We further extracted the class prototypes (Snell et al., 2017) learned by baseline TFA with the NWPU VHR-10 dataset in a few-shot setting, and visualizes the pairwise cosine similarity between them in Fig. 1. The detailed extraction procedure is given in Appendix. For instance, the similarity between the ground track field and ship is 0.05, while that between the tennis court and baseball diamond is 0.48. The classification of similar objects from different categories can easily go wrong. To address this problem, the key is to learn robust and distinctive representations for object proposals. Metric learning can achieve this by pushing apart dissimilar input samples and pulling similar ones closer in the embedding space. It has proven its effectiveness in tasks such as image retrieval (Wang et al., 2019a), classification (Luo et al., 2014; Ding and Fu, 2016; Bucher et al., 2016; Li et al., 2020c), person re-identification (Hermans et al., 2017), and face verification (Chopra et al., 2005; Wang et al., 2018). In light of this, we design a metric-learning-based approach, namely the Multi-Similarity (MS) encoding module, to better model the intra-class similarity and the inter-class difference of region proposals, thereby improving the detection performance for unseen objects.

On the other hand, catastrophic forgetting may hinder FSOD methods that follow the pretrain-transfer paradigm. Although most pretrained parameters are frozen to maintain the performance on base classes, adapting to novel data still causes a decrease in the detection accuracy on base classes. To circumvent this problem, we introduce a predictor with knowledge inheritance module to retain the Region-of-Interest (RoI) feature extractor for base classes and applying a coherence loss to guide the learning process in the fine-tuning stage. Therefore, our proposed method can better adapt to novel classes while maintaining the detection performance for base classes.

In this study, we design an original FSOD method for aerial imagery on the basis of Faster R-CNN (Ren et al., 2015) and TFA (Wang et al., 2020). Fig. 2 presents the performance of previous methods and our proposed method on the NWPU VHR-10 dataset under a 10-shot setting. Our contributions can be summarized as follows:

1. We propose an MS encoding module to boost the performance on novel categories.
2. We design a training objective, i.e., MS loss, to pull object proposal embeddings belonging to the same class together while increase the distinction between instances from different classes.
3. To circumvent the negative impact of catastrophic forgetting, we have proposed a knowledge inheritance module and design the coherence loss to achieve competitive performance on overall classes.

2. Related work

2.1. Deep learning-based object detection

Modern deep learning-based object detectors can be roughly classified into two categories: proposal-based and proposal-free methods. The landmark work, R-CNN (Girshick et al., 2014), started the trend of using CNNs in object detection. Fast R-CNN (Girshick, 2015) improved upon R-CNN by enabling end-to-end training on shared convolutional features. Ren et al. (2015) later proposed Faster R-CNN, which adopts a class-agnostic region proposal network (RPN) and achieves a much better performance in both computational speed and detection accuracy. Since then, many new methods have been developed, including RFCN (Dai et al., 2016), Mask R-CNN (He et al., 2017), and FPN (Lin et al., 2017a). Unlike proposal-based detectors, which rely on explicit object proposals, proposal-free detectors directly predict the object bounding box and classification score. Representative methods include YOLO and its variants (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020), SSD (Liu et al., 2016), and RetinaNet (Lin et al., 2017b). In this study, we follow the baseline TFA and choose Faster R-CNN with FPN as our base model.

In the aerial imagery field, it is difficult to transfer these object detectors directly because of the difference between aerial images and natural scene images (Li et al., 2020b). Some researchers modify the object detection methods and mainly focus on detecting small or cluttered objects in aerial images, such as R2-CNN (Pang et al., 2019), SSPNet (Hong et al., 2021), SCRDet (Yang et al., 2019), and SCRDet++ (Yang et al., 2022), while Polardet (Zhao et al., 2021a), S2A-Net (Han et al., 2021), and RRPN (Ma et al., 2018) aim to detect oriented targets in aerial images by adding the angle dimension. Most of these studies relied on sufficient training data to achieve an excellent performance.

2.2. Few-shot object detection

Common FSOD methods can be grouped into meta-learning (Karlinsky et al., 2019; Wang et al., 2019b; Yan et al., 2019; Han et al.,

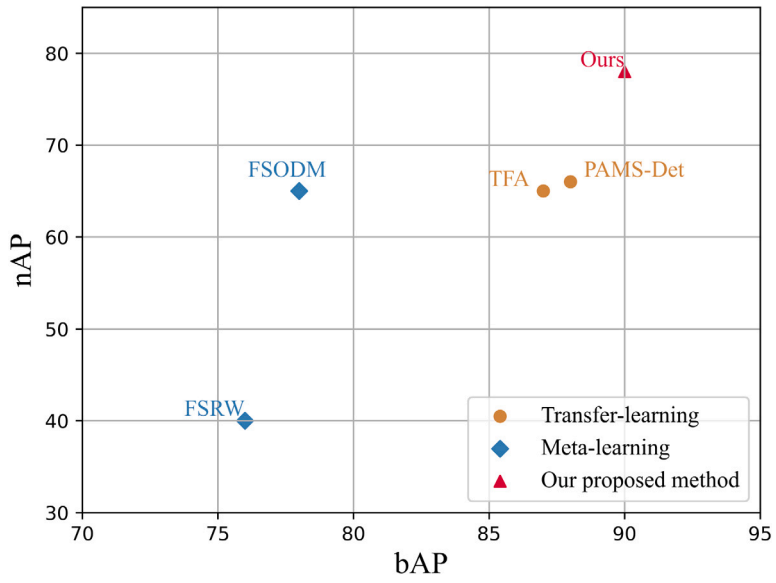


Fig. 2. Performance of existing approaches and our proposed method on the NWPU VHR-10 dataset under 10-shot setting. The average precision (AP) for base and novel categories are represented by *x*-axis and *y*-axis respectively.

2022; Kang et al., 2019) and transfer-learning methods (Wang et al., 2020; Sun et al., 2021; Fan et al., 2021; Wu et al., 2020). Meta-learning algorithms usually devise an episode-based meta-training paradigm such that the model can learn task-agnostic meta-knowledge from base classes using abundant data and adapt quickly to novel classes. Prototype matching (Yan et al., 2019; Han et al., 2022), feature reweighting (Wang et al., 2019b; Kang et al., 2019) and metric-learning (Karlinsky et al., 2019) networks are commonly used approaches in meta-detectors. Regarding transfer-learning methods, the TFA (Wang et al., 2020) pretrains the model on the base data and then refines the last few layers with novel data. This simple method outperforms previous meta-learning methods and has shows great potential for solving FSOD tasks. Subsequent studies (Sun et al., 2021; Fan et al., 2021; Wu et al., 2020) have improved the TFA by using more advanced methods. For example, FSCE (Sun et al., 2021) presents a contrastive-aware encoding module for identifying object proposals categories. Retentive R-CNN (Fan et al., 2021) proposes a parallel detector to maintain performance on base classes. MPSR (Wu et al., 2020) uses an auxiliary branch to enrich object scales by generating multi-scale positive samples and refining the prediction at different scales; the refinement process requires the manual selection of specific features. All these methods were developed for natural images.

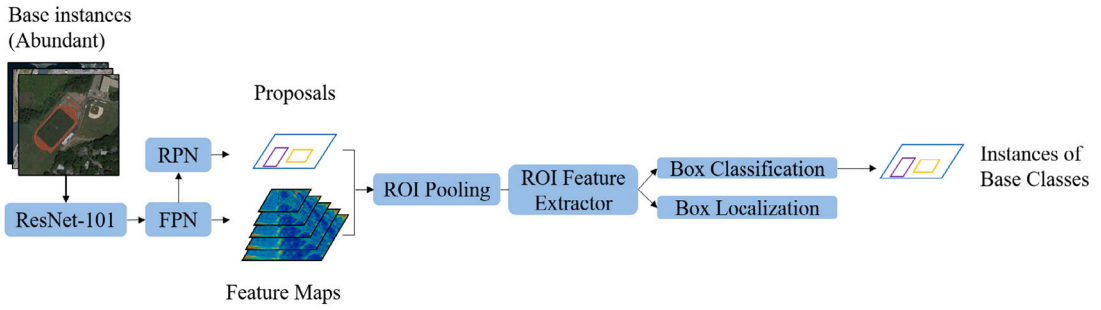
For aerial images, Li et al. (2022) introduced a meta-learning-based method that utilizes a multi-scale reweighting module to better capture class-specific information at different scales. This study represents the initial endeavor to specifically address the challenge of FSOD in aerial imagery. To detect oriented objects, Cheng et al. (2021) extended on Faster R-CNN and proposed Prototype-CNN (P-CNN) model. The prototype-guided RPN incorporated in this model can effectively detect objects with arbitrary orientations in complex backgrounds. However, meta-learning-based methods often exhibit complex designs that render them difficult to train and prone to overfitting. Xiao et al. (2021) concentrated on the relations existing at the instance-level rather than those of the entire image using a self-adaptive attention network. They employed a two-stage training paradigm and only fine-tuned the last layer; the requirements for the query and support set images still remained somewhat complex. The PAMS-Det model proposed by Zhao et al. (2021b), which is built on the TFA framework, effectively extract the multi-scale features of objects through the path aggregation module, further enhancing the detection accuracy for novel data. This research mainly focused on improving the performance on novel data, overlooking the impact of the catastrophic forgetting problem.

2.3. Deep metric learning

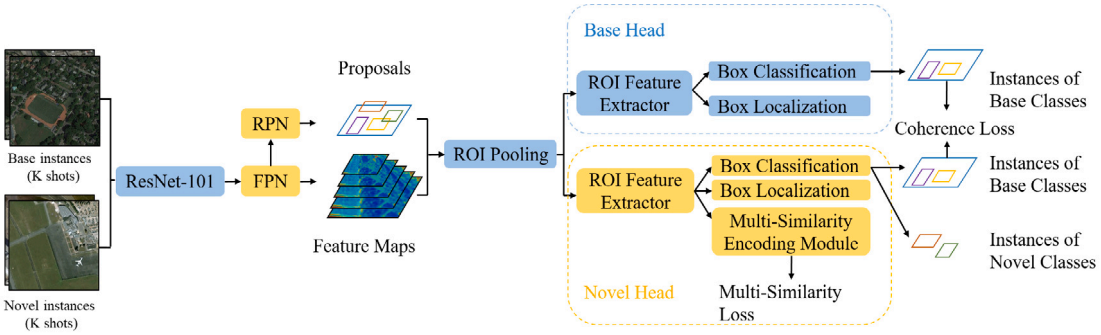
Metric-learning-based models are designed to learn semantically meaningful representations. Input samples belonging to the same category are projected close to each other, while samples from different categories are projected farther apart in the metric space. Recent studies on deep metric learning typically train samples by constructing them into pairs, triplets, or quadruplets, and use different kinds of pairwise loss functions to learn the metric. Therefore, the way of sampling pairs and weighting them is crucial to pair-based metric learning methods. For instance, contrastive loss (Hadsell et al., 2006) only considers the cosine similarity of a pair. Hoffer and Ailon (2015) and Oh Song et al. (2016) mainly focus on relative similarity which considers the similarity-relationships with other pairs. MS loss (Wang et al., 2019a) comprehensively considers all the similarity-relationships mentioned above, but it is used for image retrieval. These studies demonstrate that combining deep models with proper metric-learning-based objectives is effective in minimizing intra-class variations and maximizing inter-class variations. In this study, we design a metric-learning-based MS encoding module to better learn instance-level features and enhance the performance for novel classes.

2.4. Knowledge inheritance

Human learners can absorb the knowledge summarized by their predecessors, allowing them to adapt efficiently to new tasks. Similarly, inheriting the knowledge gained in the base training stage is helpful when adapt to novel data. However, simply transferring the parameters can cause catastrophic forgetting and performance degradation on base classes. Instead, we could inherit the semantic knowledge learned in base training stage and take the predicted score distribution as a regularization. We refer to this process as knowledge inheritance, which is inspired by knowledge distillation (Hinton et al., 2015; Li et al., 2020a). The main difference is that traditional knowledge distillation uses teacher models to provide knowledge to a lightweight student model. In the case of knowledge inheritance, due to the larger capacities of the network in the fine-tuning stage, the performance of the teacher model is no longer the limit of the student model.



(a) Base training stage



(b) Few-shot fine-tuning stage

Fig. 3. Overview of the architecture of the proposed model. (a) In the first training stage, we follow the original Faster R-CNN detection pipeline. (b) In the fine-tuning stage, the knowledge learned from the base training stage is reserved, and the multi-similarity encoding head is applied to extract discriminative representations. The combination of the proposed modules can improve the detection accuracy for novel classes and maintain the base class performance.

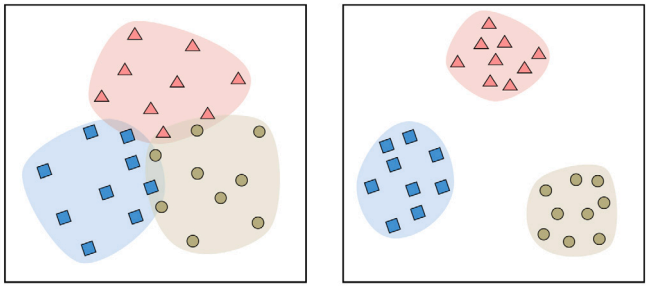
3. Methodology

This section describes the overall structure of the proposed approach. Subsequently, the MS encoding and knowledge inheritance modules are introduced.

3.1. Method overview

The entire pipeline of the proposed method is illustrated in Fig. 3. First, the backbone network takes in an image, and the RPN coarsely generates region proposals potentially containing objects. Then, the RoI head refines region proposals and generates the final predictions. Specifically, the RoI pooling layer first combines region proposals with feature maps extracted from the backbone network, then pools them to fixed sizes. The RoI feature extractor, which consists of two fully connected (FC) layers with a size of 1×1024, encodes them as RoI features. Finally, the classification and regression layers take in RoI features to generate the prediction results. We mark the RoI feature extractor and the detection head for base classes as "base head" and its parallel counterpart for novel classes as "novel head", as presented in Fig. 3(b).

We adopt a classical two-stage pretrain-transfer scheme. Following previous studies (Kang et al., 2019; Li et al., 2022; Zhao et al., 2021b; Wolf et al., 2021), the categories of a dataset are divided into novel classes C_{novel} and base classes C_{base}, with D_{novel} and D_{base} denoting their sub-datasets, respectively. D_{base} contains abundant base data for training, whereas D_{novel} is a small balanced training dataset with K samples of C_{base} and C_{novel}. In the first stage, we train the vanilla Faster R-CNN on D_{base}, to obtaining a base model f_{base}. Then, a novel model f_{novel} is obtained by fine-tuning f_{base} on D_{novel}. The backbone network is fixed, along with the base head, to act as a regularization for the novel head. The proposed MS loss and coherence loss are jointly optimized with the original Faster R-CNN training objectives to improve the overall performance.



(a) Object proposal embeddings learned by softmax

(b) Object proposal embeddings learned by MS encoding module

Fig. 4. Conceptualization of the proposed MS encoding module for object region proposals. We design a proposal module branch and adopt a score function to discriminatively learn a similarity metric.

3.2. Multi-similarity encoding module

To establish distinctive and robust embeddings for object proposals, we incorporate the MS encoding module into the novel head alongside the original output layers. Since the RoI feature extractor consists of two FC layers with rectified linear unit (ReLU) (Vinod and Hinton, 2010) as the activation function, the RoI features are truncated at zero. To mitigate the potential influence caused by the non-negativity constraint (Dubey et al., 2022), we attach an FC layer to project the RoI features into a 128-dimensional embedding space, ensuring reliable metric measurement. Then, we measure the metric distance between object proposal embeddings and optimize it to maximize the variance between those with different labels and minimize the distinction between those belonging to the same category. This results in a larger margin and improved decision boundary, benefiting classification and regression tasks, as shown in Fig. 4. An illustration of the proposed MS encoding module is presented in Fig. 5.

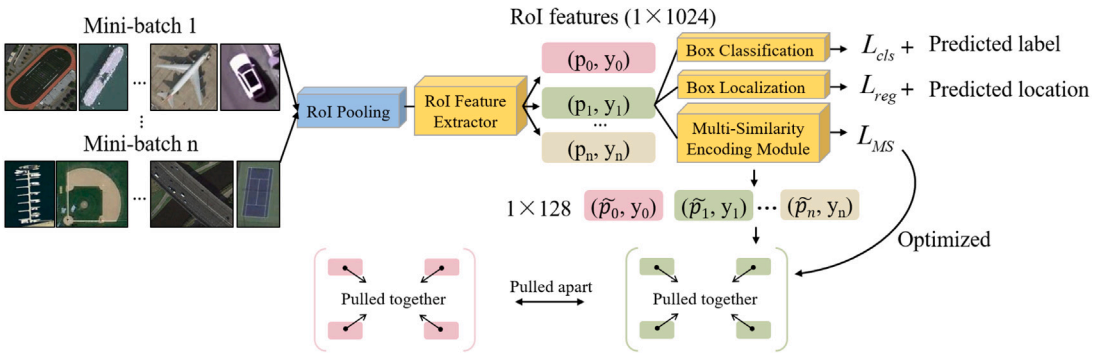


Fig. 5. Illustration of the proposed MS encoding module. The MS encoding module induces the RoI feature extractor to capture useful information and extract distinctive proposal embeddings. By optimizing the MS loss function, embeddings with the same label are pulled together, while ones from different classes are pushed away in the embedding space. In this way, the margin around each cluster is expanded thus leading to a more defined decision boundary.

Inspired by Wang et al. (2019a) in image retrieval, we have designed our MS loss function that is tailored for object detection tasks. Specifically, given a batch of m RoI box features, $\{x_i, u_i, y_i\}_{i=1}^{m}$, where x_i is the RoI feature encoded by the MS module for the i-th object proposal, u_i denotes its Intersection-over-Union (IoU) score with the corresponding ground truth box, and y_i represents the ground truth label. S_{ij} denotes the cosine similarity between the ith and jth proposals in the projected unit hypersphere:

$$S_{ij} = \frac{x_i}{\|x_i\|} \cdot \frac{x_j}{\|x_j\|} \tag{1}$$

The MS loss is formulated as:

$$L_{MS} = \frac{1}{m}\sum_{i=1}^{m}\left\{ \mathrm{II}\left(u_i \geq \varphi\right)\frac{1}{\alpha}\log\left[1+\sum_{y_i=y_j} e^{-\alpha(S_{ij}-\lambda)}\right] + \frac{1}{\beta}\log\left[1+\sum_{y_i\neq y_j} e^{\beta(S_{ij}-\lambda)}\right]\right\}, \tag{2}$$

where α, β, and λ are fixed hyperparameters as in Wang et al. (2019a). $\frac{x_i}{\|x_i\|}$ denotes the normalized features and $\mathrm{II}\left(u_i \geq \varphi\right)$ is the indicator function. In Eq. (2), we propose applying an IoU threshold φ to ensure that the object proposals are sampled near the object and reduce irrelevant semantics it might contain. We empirically find $\varphi = 0.7$ is good enough for the MS encoding module to be trained with the most centered region proposals. The first term of Eq. (2) minimizes the similar pairs $\left(u_i \geq \varphi\right)$ distances, whereas the second term penalizes the dissimilar pairs distances for being too close.

3.3. Knowledge inheritance module

Owing to the catastrophic forgetting effect, f_{novel} tends to degrade the performance of base classes. We aim to learn a detection model $f(\cdot)$ for both C_{base} and C_{novel} from D_{novel} without hindering the performance of base classes. To achieve this, the base head of f_{base} is retained with its parameters frozen. Then, the novel head, initialized with the parameters of f_{base}, is introduced and adjusted during the fine-tuning stage. However, simply separating the predictions of the base and novel classes can fail to establish interactions between the two branches and potentially lead to a suboptimal system. Since f_{base} is trained using sufficient data, its prediction of objects in C_{base} is considered as a reliable result. To inherit knowledge from f_{base}, we design an auxiliary coherence loss to regularize f_{novel} and align its classification probability distributions with those of f_{base} for the base class objects. The total number of base classes is denoted as N_{base}. For a given input instance, the classification probabilities predicted by f_{base} and f_{novel} for class c (where $c \in C_{\text{base}}$) are represented as p^c_{base} and p^c_{novel}, respectively. The coherence loss is defined as:

$$L_{\text{coherence}} = \sum_{c=1}^{N_{\text{base}}} \tilde{p}^c_{\text{base}} \log\left(\frac{\tilde{p}^c_{\text{base}}}{\tilde{p}^c_{\text{novel}}}\right), \tag{3}$$

where

$$\begin{cases} \tilde{p}^c_{novel} = \dfrac{p^c_{novel}}{\sum_{i=1}^{N_{base}} p^i_{novel}} \\ \tilde{p}^c_{base} = \dfrac{p^c_{base}}{\sum_{i=1}^{N_{base}} p^i_{base}} \end{cases}.$$

Note that $\tilde{p}^c_{novel}$ and p^c_{novel} are scalars, and the former refers to the normalized probability of the latter among base classes. This normalization ensures that the coherence loss calculation is based on a probability distribution that only considers base classes. The purpose is to encourage the predictions of f_{novel} for base classes to align with those of f_{base}. Thus, our proposed approach facilitates the prediction of novel objects and also maintains the performance on base classes.

3.4. Training objectives

In the first training stage, we adopt a standard Faster R-CNN loss function (Ren et al., 2015) as the training objective. It contains three parts: loss L_{rpn} for generating high-quality region proposals, loss L_{cls} for classifying the bounding boxes, and loss L_{reg} for predicting the bounding box regression offsets. In the fine-tuning stage, we froze the backbone feature extractor as well as the base head and train the remaining part using the MS loss and coherence loss along with the Faster R-CNN loss as follows:

$$L_{\text{total}} = L_{\text{rpn}} + L_{\text{cls}} + L_{\text{reg}} + \lambda_1 L_{\text{MS}} + \lambda_2 L_{\text{coherence}}, \tag{4}$$

where the coefficients λ_1 and λ_2 are set to 0.5 to balance the scale of these losses. In the inference stage, the object proposals generated by the RPN are only fed into the novel head for final prediction.

4. Experiments

In this section, we first describe the experimental settings. Then, we compare our proposed methods with state-of-the-art FSOD methods on aerial images. Finally, we conduct extensive ablation experiments to demonstrate the effectiveness of each proposed component.

4.1. Few-shot detection benchmarks

All experiments are performed on both the NWPU VHR-10 and DIOR benchmarks. We follow the same splits and sampling method of base and novel categories utilized in previous works (Li et al., 2022; Zhao et al., 2021b) to ensure a fair comparison.

(1) NWPU VHR-10 Dataset. This is a 10-class object detection geospatial dataset with a spatial resolution ranging from 0.5 to 2 m. It consists of 800 high-resolution aerial images and 650 of which contain objects belonging to annotated categories. These ten classes of objects are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle.

Table 1
Comparisons of FSOD performance (AP50) on novel classes of the NWPU VHR-10 dataset.

Method	TFA			FSODM			FSRW			PAMS-Det			Ours		
	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot	3-shot	5-shot	10-shot
airplane	0.12	0.51	0.60	0.15	**0.58**	0.60	0.13	0.24	0.20	0.21	0.55	0.61	**0.41**	0.53	**0.77**
baseball diamond	0.61	0.78	0.85	0.57	0.84	0.88	0.12	0.39	0.74	**0.76**	**0.88**	0.88	0.66	0.81	**0.90**
tennis court	0.13	0.19	0.49	0.25	0.16	0.48	0.11	0.11	0.26	0.16	0.20	0.50	**0.29**	**0.49**	**0.68**
mean	0.29	0.49	0.65	0.32	0.53	0.65	0.12	0.24	0.40	0.37	0.55	0.66	**0.45**	**0.61**	**0.78**

Table 2
Comparisons of FSOD performance (AP50) on novel classes of the DIOR dataset.

Method	TFA			FSODM			FSRW			PAMS-Det			Ours		
	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot	5-shot	10-shot	20-shot
airplane	0.13	0.17	0.24	0.09	0.16	0.22	0.09	0.15	0.19	0.14	0.17	0.25	**0.15**	**0.32**	**0.59**
baseball field	0.51	0.53	0.56	0.27	0.46	0.50	0.33	0.45	0.52	**0.54**	**0.55**	0.58	0.46	0.53	**0.60**
tennis court	0.24	0.41	0.50	**0.57**	**0.60**	**0.66**	0.47	0.54	0.55	0.24	0.41	0.50	0.48	0.48	0.55
train station	0.13	0.15	0.21	0.11	0.14	0.16	0.09	0.07	0.18	0.17	**0.17**	**0.23**	**0.17**	0.16	0.18
wind mill	0.25	0.30	0.33	0.19	0.24	0.29	0.13	0.18	0.26	**0.31**	**0.34**	**0.36**	0.17	0.20	0.26
mean	0.25	0.31	0.37	0.25	0.32	0.36	0.22	0.28	0.34	0.28	0.33	0.38	**0.28**	**0.34**	**0.40**

Table 3
Comparisons of FSOD performance (AP50) on base classes of the NWPU VHR-10 dataset.

	TFA	FSODM	FSRW	PAMS-Det	Ours
ship	0.86	0.72	0.77	0.88	**0.90**
storage tank	0.89	0.71	0.80	0.89	**0.90**
basketball court	0.89	0.72	0.51	0.90	**0.91**
ground track field	0.99	0.91	0.94	0.99	**0.99**
harbor	0.84	0.87	0.86	0.84	**0.90**
bridge	0.78	0.76	0.77	0.80	**0.84**
vehicle	0.87	0.76	0.68	0.89	**0.90**
mean	0.87	0.78	0.76	0.88	**0.90**

Table 4
Comparisons of FSOD performance (AP50) on base classes of the DIOR dataset.

	TFA	FSODM	FSRW	PAMS-Det	Ours
airport	0.76	0.63	0.59	**0.78**	0.76
basketball court	0.78	0.80	0.74	0.79	**0.81**
bridge	0.52	0.32	0.29	**0.52**	0.47
chimney	0.66	0.72	0.70	0.69	**0.78**
dam	0.54	0.45	0.52	0.55	**0.64**
expressway service area	0.66	0.63	0.63	0.67	**0.78**
expressway toll station	0.60	0.60	0.48	0.62	**0.73**
golf course	0.79	0.61	0.61	**0.81**	0.78
ground track field	0.77	0.61	0.54	0.78	**0.80**
harbor	0.50	0.43	**0.52**	0.50	0.45
overpass	0.50	0.46	0.49	0.51	**0.59**
ship	0.66	0.50	0.33	0.67	**0.72**
stadium	0.75	0.45	0.52	**0.76**	0.74
storage tank	0.55	0.43	0.26	0.57	**0.71**
vehicle	0.52	0.39	0.29	0.54	**0.56**
mean	0.63	0.54	0.50	0.65	**0.69**

(2) DIOR Dataset. This is a challenging 20-class object detection dataset which contains 23463 aerial images with 192,472 object instances. Each images has a size of 800 × 800 px, and the spatial resolution ranges from 0.5 to 30 m. The 20 object classes are airplane, basketball court, airport, baseball field, bridge, chimney, dam, expressway service area, expressway toll station, golf course, vehicle, ground track field, harbor, ship, stadium, storage tank, overpass, tennis court, train station, and wind mill.

As presented in Table 1, the overall 10 classes in the NWPU VHR-10 dataset are separated into 7 base categories and 3 novel categories. Similarly, in the DIOR dataset, out of the 20 categories, 5 are designated as novel categories, and the other 15 categories remain as base categories. All base classes are available, and K-shot of samples are randomly selected from novel classes for few-shot training. Following existing works (Li et al., 2022; Zhao et al., 2021b), we set K = 3, 5, and 10 for the NWPU VHR-10 dataset and K = 5, 10, and 20 for the DIOR dataset. To evaluate the performance more accurately, we adopt the mean average precision (mAP) metric and consider detections with an IoU > 0.5 to a ground truth annotation as correct. The results are evaluated on a test dataset that includes both base and novel categories, since the overall performance is our major concern.

4.2. Implementation details

We follow the baseline TFA and adopt the ResNet-101 (He et al., 2016) and FPN (Lin et al., 2017a) as the backbone network. For each level of pyramid features (P_2 to P_6), we define one squared anchor per location with areas of $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ pixels. The solver is the standard SGD optimizer with weight decay 0.0001 and momentum 0.9. During the base training stage, we set the initial learning rate at 0.001, and the learning rate was decreased by 0.1 at epochs 12 and 18. In the fine-tuning stage, we set the initial learning rate at 0.0005, and the learning rate was decreased by 0.1 at 4000 iterations. We train 24 epochs in base training and finetune till full convergence. We adopt a learning rate warm-up for 100 iterations. Due to the memory limitation, we set batch size at 4 for the NWPU VHR-10 dataset and 8 for the DIOR dataset. Finally, the hyperparameters in Eq. (2) are: $\alpha = 2, \lambda = 1, \beta = 50; \lambda_1$ and λ_2 in Eq. (4) are set to 0.5. The training and testing are executed using a NVidia GeForce RTX 3090Ti GPU with 24-GB memory.

4.3. Comparison of results

We compared our proposed method with the prevalent fine-tuning based method TFA (Wang et al., 2020), which serves as the baseline method, the meta-learning-based method FSRW (Kang et al., 2019), and the FSODM (Li et al., 2022) on aerial images, as well as the state-of-the-art fine-tuning-based method PAMS-Det (Zhao et al., 2021b). For the backbone network, the FSRW adopts the YOLOv2 network, and the FSODM adopts the YOLOv3 network. The PAMS-Det designs a backbone similar to the ResNet and FPN. We did not implement additional data augmentation techniques in our proposed method for a fair comparison.

The FSOD performance of the proposed approach on the NWPU VHR-10 dataset is presented in Table 1. It is evident that our proposed model outperforms the baseline method TFA, FSODM, and PAMS-Det on both base and novel classes. Specifically, for novel classes, our method obtains an mAP that is 13% higher in the 3-shot setting, 8% higher in the 5-shot setting, and 13% higher in the 10-shot setting than FSODM. Furthermore, our method achieves an mAP that is 8% higher in the 3-shot setting, 6% higher in the 5-shot setting, and 12% higher in the 10-shot setting than PAMS-Det. Regarding base classes, our method

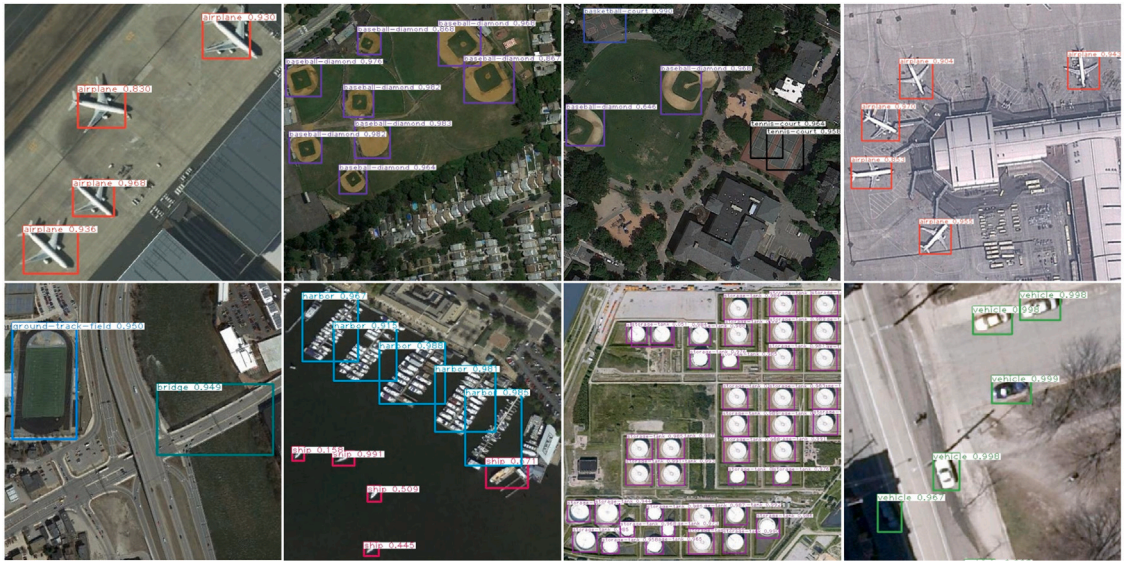


Fig. 6. Examples of the 10-shot detection results of the proposed method on the NWPU VHR-10 dataset. The prediction results for the novel and base classes are shown in the first and second rows, respectively.

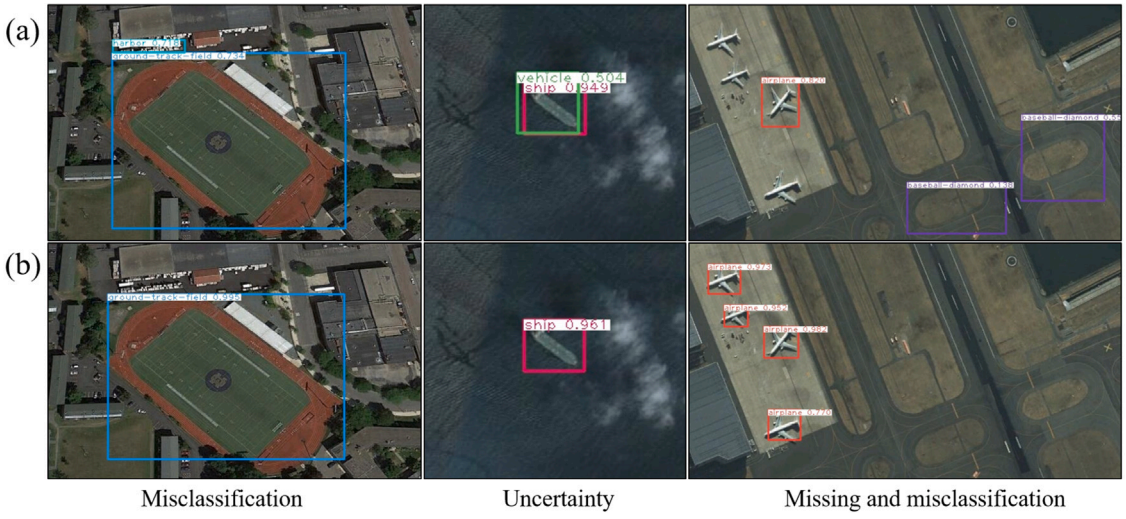


Fig. 7. Visualization of failure cases avoided by the proposed method in the 10-shot setting on the NWPU VHR-10 dataset. (a) The first row presents the object detection results of the baseline method TFA. (b) The second row presents the object detection results of the proposed method.

achieves the best performance compared to the other three methods as shown in Table 3. In concrete terms, our method achieves an mAP 12% that is higher on base classes than FSODM, and 2% higher than PAMS-Det. The qualitative results of our proposed model on the NWPU VHR-10 dataset are given in Fig. 6.

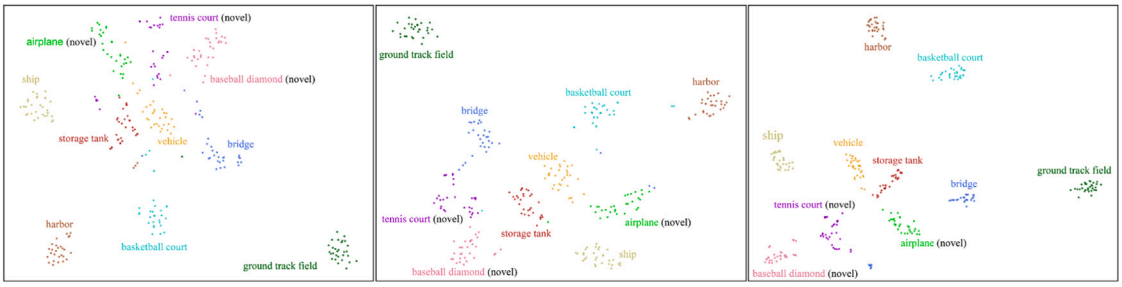
Regarding the DIOR dataset, our method achieves a better performance than FSODM and PAMS-Det. As shown in Table 2, our method achieves 28%, 34%, and 40% mAPs for novel classes of 5-shot, 10-shot, and 20-shot, respectively; and the performance of most novel classes is improved. We also compare the detection performance for base classes: the mAP of our proposed method is 15% higher than that of FSODM and 4% higher than PAMS-Det as shown in Table 4.

Our proposed method achieves a satisfactory performance on novel classes without compromising the detection accuracy of base class objects. Additionally, both TFA and PAMS-Det achieved higher mAP values on base classes compared to FSODM and FSRW. This suggests that fine-tuning-based methods could circumvent the performance degradation of base classes better than the meta-learning-based methods could. Fig. 7 shows examples of the failure cases of the baseline method, such as missing detections for novel instances, ambiguous classifications for detected objects, and misclassifications. In contrast, our proposed method performed well in addressing these challenges.

4.4. Ablation study

In this section, we conducted ablation experiments on the NWPU VHR-10 dataset to demonstrate the effectiveness of our proposed approach in the 10-shot scenario. We also examined the impact of hyperparameters in MS encoding module; the results are presented in Table 6.

Ablation study for the two components proposed in our method. We apply the MS encoding module and knowledge inheritance module to the baseline TFA step-by-step and report the performance. As shown in Table 5, the MS encoding module can achieve a 6% improvement on novel classes compared to the baseline method. This demonstrates that the MS loss can reduce misclassification by guiding the model to learn distinctive proposal embeddings. However, solely employing the MS encoding module does not prevent the detector



(a) Object proposal embeddings learned without MS loss or coherence loss

(b) Object proposal embeddings learned with coherence loss only

(c) Object proposal embeddings learned with MS loss and coherence loss

Fig. 8. Visualization of the object proposal embeddings learned with and without the proposed losses using t-SNE. This figure presents proposal embeddings of 30 instances randomly selected from each class in the NWPU VHR-10 dataset in a 10-shot setting. After applying the MS encoding module, the clusters became tighter, and the margin around clusters widened. The proposed modules enable the model to learn distinctive embeddings to separate the ten classes effectively.

Table 5
Ablative performance (mAP50) of key components proposed in our method on the NWPU VHR-10 dataset in 10-shot setting.

	MS loss	Coherence loss	bAP	nAp	mAp
TFA	✗	✗	0.87	0.65	0.80
Our proposed methods	✓	✗	0.85	0.71	0.81
	✗	✓	0.89	0.68	0.83
	✓	✓	0.90	0.78	0.87

Table 6
Impact of MS encoding module dimension and hyperparameter λ in the MS encoding module.

MS encoding module dimension	$\lambda = 0.2$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2$
1×64	0.83	0.84	0.86	0.83	0.81
1×128	0.85	0.86	0.87	0.85	0.82
1×256	0.85	0.85	0.86	0.85	0.82

from catastrophic forgetting. We observed a decrease in performance for base classes, dropping from 87% to 85%. On the other hand, by applying the coherence loss alone, we achieved a 2% performance improvement for base classes and a 3% improvement for novel classes. To achieve the best results, we combined the MS encoding module with the knowledge inheritance module. This combination led to a 3% performance improvement for base classes and an impressive 13% improvement for novel classes. The entire system benefits from these two proposed components.

Impact of the head dimension and hyperparameter λ in the MS encoding module. As the primary RoI feature extractor contains ReLU layers that truncate features at zero, we encode the RoI features with an MS encoding module to a 128-dimensional vector to meaningfully measure the metrics. In this section, we conducted experiments to explore the impact of the MS encoding module dimension and the hyperparameter λ. We observed that setting λ to 1 and the dimension of the MS encoding module to 1×128 yielded better results. Moreover, we noticed that a slight improvement in detection performance as the dimension increased from 1×64 to 1×128. However, the performance of dimension 1×128 is either similar or slightly higher than the performance of dimension 1×256 across all values of λ. Therefore, we concluded that the difference in performance between the two dimensions is relatively small and may not be significant in practice. Based on these findings, we set the dimension at 1×128.

Visualization and analysis. Fig. 8 presents the t-SNE (Van der Maaten and Hinton, 2008) visualization of proposal embeddings trained with and without the proposed components on the same set of samples. The baseline method can learn a well-separated decision boundary for base classes with abundant samples, but not for novel classes. The proposal embeddings from base classes benefit significantly from the knowledge inheritance module. For example, a comparison of Fig. 8.(a) and (b) reveals that the proposal embeddings of vehicles have a much more defined boundary than those of storage tanks and bridges. This indicates that the coherence loss helps the model better preserve the knowledge learned from base training stage. Fig. 8(c) also reveals that after the application of the MS encoding module, the proposal embeddings gather together and the instance clusters become much tighter, resulting in the margin around clusters broaden. This demonstrates the effectiveness of the MS encoding module in increasing the compactness of object proposal embeddings from the same category and the distinctiveness of those from different categories. Learning embeddings with compactness and distinctiveness makes it easier for the detector to correctly classify detected objects. This is in alignment with the results of the ablation studies and confirm the effectiveness of the proposed methods.

5. Conclusion

In this work, we propose a novel FSOD approach in aerial imagery using deep metric learning and knowledge inheritance. Our proposed method incorporates the MS encoding module, which enables the model to learn robust and distinctive embeddings for object proposals, resulting in enhanced performance on novel categories of aerial imagery. To overcome the weakness of catastrophic forgetting, we build a knowledge inheritance module to better preserve the knowledge learned from the base training stage. By building on top of generalized feature representations, our proposed method achieves satisfactory results on novel categories while remaining competitive on base categories. Experiments conducted on well-established benchmarks demonstrate that our model reaching the state-of-the-art overall performance among all data settings; ablation studies further validate the effectiveness of our proposed method. The FSOD task has great industrial potential, as its successes would enable a variety of visual tasks to gain the generalization ability on unseen objects without heavily consuming labor annotation. We believe that our work can contribute to further research in addressing FSOD challenges for aerial imagery.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Appendix

A. Class prototypes

We followed the class prototype extraction method described in Snell et al. (2017). The concrete procedure is described as follows: First, we split the NWPU VHR-10 dataset into novel classes (airplane, baseball diamond, and tennis court) and base classes (the remaining 7 classes). Next, we train the baseline method TFA (Wang et al., 2020) with abundant base class data, and freeze the backbone parameters for transferring. Then, we fine-tune the network with a small balanced dataset (randomly picked 10 samples from each novel classes and base classes). Finally in the class prototype extraction procedure, we put 30 instances of each class (both novel and base classes) into the network and collect the representations learned by the penultimate layer of the network. Few-shot class prototypes are computed as the mean vector of the representations in the embedding space for each class. We denote the set of N labeled instances as $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where each $x_i \in \mathbb{R}^{\mathbb{D}}$ is the D-dimensional feature vector of an instance, and $y \in \{1, 2, \ldots, K\}$ is the corresponding label. S_k denotes the set of instances labeled with class k. The class prototype $c_k \in \mathbb{R}^{\mathbb{M}}$ of each class is computed through the network function $f_\varphi : \mathbb{R}^D \to \mathbb{R}^M$ with trainable parameters φ. So we can calculate the class prototype of class k as:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\varphi(x_i). \tag{5}$$

By extracting the prototype of each class, we are able to calculate the cosine similarity of each pair of classes.

References

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Bucher, M., Herbin, S., Jurie, F., 2016. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: European Conference on Computer Vision. Springer, pp. 730–746.

Chen, H., Wang, Y., Wang, G., Qiao, Y., 2018. Lstd: A low-shot transfer detector for object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS J. Photogramm. Remote Sens. 117, 11–28.

Cheng, G., Yan, B., Shi, P., Li, K., Yao, X., Guo, L., Han, J., 2021. Prototype-CNN for few-shot object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–10.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1. CVPR'05, IEEE, pp. 539–546.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. Adv. Neural Inf. Process. Syst. 29.

Ding, Z., Fu, Y., 2016. Robust transfer metric learning for image classification. IEEE Trans. Image Process. 26 (2), 660–670.

Dubey, S.R., Singh, S.K., Chaudhuri, B.B., 2022. Activation functions in deep learning: A comprehensive survey and benchmark. Neurocomputing.

Fan, Z., Ma, Y., Li, Z., Sun, J., 2021. Generalized few-shot object detection without forgetting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4527–4536.

French, R.M., 1999. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences 3 (4), 128–135.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. CVPR'06, IEEE, pp. 1735–1742.

Han, J., Ding, J., Li, J., Xia, G.S., 2021. Align deep features for oriented object detection. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Han, G., Huang, S., Ma, J., He, Y., Chang, S.F., 2022. Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. pp. 780–789.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737.

Hinton, G., Oriol, V., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hoffer, E., Ailon, N., 2015. Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition. Springer, pp. 84–92.

Hong, M., Li, S., Yang, Y., Zhu, F., Zhao, Q., Lu, L., 2021. SSPNet: Scale selection pyramid network for tiny person detection from UAV images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T., 2019. Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8420–8429.

Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M., 2019. Repmet: Representative-based metric learning for classification and few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5197–5206.

Kim, J.H., Ryu, S., Jeong, J., So, D., Ban, H.J., Hong, S., 2020. Impact of satellite sounding data on virtual visible imagery generation using conditional generative adversarial network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4532–4541.

Li, X., Deng, J., Fang, Y., 2022. Few-shot object detection on remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–14. http://dx.doi.org/10.1109/TGRS.2021.3051383.

Li, Y., Francis, E.T., Li, G., Wang, T., Shifeng, J., 2020a. Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020b. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. 159, 296–307.

Li, X., Yu, L., Fu, C.W., Fang, M., Heng, P.A., 2020c. Revisiting metric learning for few-shot image classification. Neurocomputing 406, 49–58.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: European Conference on Computer Vision. Springer, pp. 21–37.

Liu, T., Zhang, L., Wang, Y., Guan, J., Fu, Y., Zhou, S., 2022. An empirical study and comparison of recent few-shot object detection algorithms. arXiv preprint arXiv:2203.14205.

Luo, Y., Liu, T., Tao, D., Xu, C., 2014. Decomposition-based transfer distance metric learning for image classification. IEEE Trans. Image Process. 23 (9), 3789–3801.

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X., 2018. Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. Multimed. 20 (11), 3111–3122.

Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S., 2016. Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4004–4012.

Pang, J., Li, C., Shi, J., Xu, Z., Feng, H., 2019. R2-CNN: Fast tiny object detection in large-scale remote sensing images. arXiv preprint arXiv:1902.06042.

Qiu, C., Tong, X., Schmitt, M., Bechtel, B., Zhu, X.X., 2020. Multilevel feature fusion-based CNN for local climate zone classification from sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 2793–2806.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7263–7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Shi, L., Huang, X., Zhong, T., Taubenböck, H., 2019. Mapping plastic greenhouses using spectral metrics derived from GaoFen-2 satellite data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 49–59.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Snell, J., Kevin, S., Richard, Z., 2017. Prototypical networks for few-shot learning. Adv. Neural Inf. Process. Syst. 30.

Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C., 2021. Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7352–7362.

Tang, Z., Wang, H., Li, X., Li, X., Cai, W., Han, C., 2020. An object-based approach for mapping crop coverage using multiscale weighted and machine learning methods. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1700–1713.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Vinod, N., Hinton, G., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning. ICML-10.

Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R., 2019a. Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5022–5030.

Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F., 2020. Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957.

Wang, Y.-X., Ramanan, D., Hebert, M., 2019b. Meta-learning to detect rare objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9925–9934.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274.

Wolf, S., Meier, J., Sommer, L., Beyerer, J., 2021. Double head predictor based few-shot object detection for aerial imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 721–731.

Wu, J., Liu, S., Huang, D., Wang, Y., 2020. Multi-scale positive sample refinement for few-shot object detection. In: European Conference on Computer Vision. Springer, pp. 456–472.

Xiao, Z., Qi, J., Xue, W., Zhong, P., 2021. Few-shot object detection with self-adaptive attention network for remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 4854–4865.

Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L., 2019. Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9577–9586.

Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., He, T., 2022. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE Trans. Pattern Anal. Mach. Intell.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K., 2019. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8232–8241.

Zhao, P., Qu, Z., Bu, Y., Tan, W., Guan, Q., 2021a. Polardet: A fast, more precise detector for rotated target in aerial images. Int. J. Remote Sens. 42 (15), 5831–5861.

Zhao, Z., Tang, P., Zhao, L., Zhang, Z., 2021b. Few-shot object detection of remote sensing images via two-stage fine-tuning. IEEE Geosci. Remote Sens. Lett. 19, 1–5.